

Empirical Bayes improvement of Kalman filter type of estimators

Eitan Greenshtein

Israel Census Bureau of Statistics; e-mail: eitan.greenshtein@gmail.com

Ariel Mansura,

Bank of Israel; e-mail: ariel.mansura@boi.org.il

Ya'acov Ritov

The Hebrew University of Jerusalem; e-mail: yaacov.ritov@gmail.com

Abstract: We consider the problem of estimating the means μ_i of n random variables $Y_i \sim N(\mu_i, 1)$, $i = 1, \dots, n$. Assuming some structure on the μ process, e.g., a state space model, one may use a summary statistics for the contribution of the rest of the observations to the estimation of μ_i . The most important example for this is the Kalman filter. We introduce a non-linear improvement of the standard weighted average of the given summary statistics and Y_i itself, using empirical Bayes methods. The improvement is obtained under mild assumptions. It is strict when the process that governs the states μ_1, \dots, μ_n is not a linear Gaussian state-space model. We consider both the sequential and the retrospective estimation problems.

1. Introduction and Preliminaries

We consider the estimation under squared error loss of a vector μ_1, \dots, μ_n observed with Gaussian additive error: $Y_i = \mu_i + \varepsilon_i$, $i = 1, \dots, n$, where $\varepsilon_1, \dots, \varepsilon_n$ are i.i.d. $N(0, 1)$. It is natural in our applications to consider the index i as denoting time, and regard μ_1, \dots, μ_n as a realization of a stochastic process. We analyze, accordingly, two main setups. In the first, the estimation is done retrospectively, after all of Y_1, \dots, Y_n are observed. The second case is of sequential estimation, where μ_i should be estimated at time i , after observing Y_1, \dots, Y_i . Let \mathcal{D}_i be the data set based on which μ_i is estimated, excluding the i th observation itself. That is, $\mathcal{D}_i = \{1, \dots, i-1\}$ in the sequential case, and $\mathcal{D}_i = \{j : 1 \leq j \leq n, j \neq i\}$ when the estimation is retrospective.

We could consider a more general situation in which the observations are (Y_i, \mathbf{X}_i) , $i = 1, \dots, n$, where the \mathbf{X}_i s are observed covariates and

$$\hat{\mu}_i = \sum_{j \in \mathcal{D}_i \cup \{i\}} \left(\beta_{ij} Y_j + \beta_{ij}^x \mathbf{X}_j \right).$$

However, to simplify the presentation, we discuss only the situation without observed covariates:

$$\hat{\mu}_i = \sum_{j \in \mathcal{D}_i \cup \{i\}} \beta_{ij} Y_j. \quad (1)$$

When μ_1, μ_2, \dots are a realization of a Gaussian process, the optimal estimator for μ_i based on the data set $\mathcal{D}_i \cup \{i\}$ is indeed linear, and is given by the Kalman filter (KF). However, in more general state space models, and certainly when the model is misspecified, the Kalman filter, or any other linear scheme, are not optimal. Yet, they may be taken as a reasonable starting point for the construction of a better estimator. We consider in this paper an empirical Bayes improvement of a given linear filter which is nonparametric and does not depend on structural assumptions.

The linear estimator $\hat{\mu}_i$ in (1) can be considered as a weighted average of two components, Y_i , and an estimator $\tilde{\mu}_i$ based on all the observations available at time i excluding the i th observations itself:

$$\tilde{\mu}_i = \sum_{j \in \mathcal{D}_i} \tilde{\beta}_{ij} Y_j.$$

In the Gaussian case, $\tilde{\mu}_i$ and $\hat{\mu}_i$ are typically the sufficient statistics for μ_i given the data in \mathcal{D}_i and $\mathcal{D}_i \cup \{i\}$ respectively. In the sequential Gaussian case the estimator $\tilde{\mu}_i$ is called the optimal one step ahead predictor of μ_i while $\hat{\mu}_i$ is the KF estimator of μ_i , $i = 1, \dots, n$. For background about the KF, state-space models, and general facts about time series see, e.g., Brockwell and Davis (1991). We will hardly use that theory in the following development, since we aim for results that are true regardless on whether various common state-space assumptions hold. In the sequel, when we want to emphasize that $\tilde{\mu}_i$ and $\hat{\mu}_i$ are the standard KF estimators we will write $\tilde{\mu}_i^K$ and $\hat{\mu}_i^K$, but the following derivation is for a general pair $\tilde{\mu}_i$ and $\hat{\mu}_i$.

Our goal in this paper is to use $\tilde{\mu}_i$ as a basis for the construction of an estimator which improves upon $\hat{\mu}_i$. In fact, we try to find the best estimator of the form:

$$\hat{\mu}_{ig} = \tilde{\mu}_i + g(Y_i - \tilde{\mu}_i), \quad i = 1, \dots, n. \quad (2)$$

Let

$$\delta \equiv \arg \min_g E \sum_{i=1}^n (\hat{\mu}_{ig} - \mu_i)^2 \quad (3)$$

Thus, we use a simple coordinate-wise function, as was introduced by Robbins (1951) in the context of compound decision:

Definition 1 A function $\mathbf{f} : \mathbb{R}^n \rightarrow \mathbb{R}^n$ is called *simple coordinate-wise function*, if it has the representation $\mathbf{f}(X_1, \dots, X_n) = (f(X_1), \dots, f(X_n))$ for some $f : \mathbb{R} \rightarrow \mathbb{R}$.

Our improvement, denoted $\delta(\cdot)$, is a simple coordinate-wise function of $(\mathbf{Y} - \tilde{\boldsymbol{\mu}})$. In the theory of compound decision and empirical Bayes, the search for an optimal simple-coordinate-wise decision function is central. We elaborate in the next section. The improved estimator $\mu_{i\delta}$ is denoted μ_i^I , and in vector notations we write in short

$$\boldsymbol{\mu}^I = \tilde{\boldsymbol{\mu}} + \boldsymbol{\delta}.$$

1.1. Empirical Bayes and non-exchangeable observations

The ideas of empirical Bayes (EB) and compound decision (CD) procedures were developed by Robbins (1951, 1955, 1964), see the review papers of Copas (1969) and Zhang (2003), and the paper of Greenshtein and Ritov (2008) for results relating compound decision, simple coordinate-wise decision functions and permutational invariant decision functions.

The classical EB/CD theory is restricted to independent exchangeable observations and to permutation invariant procedures, and in particular it excludes the utilization of explanatory variables. Fay and Herriot (1979) suggested a way to extend the ideas of parametric EB (i.e. linear decision functions corresponding to Gaussian measurement and prior) to handle covariates. Recently, there is an effort to extend the EB ideas, so they may be incorporated in the presence of covariates also in the context of non-parametric EB, see, Jiang and Zhang (2010), Cohen et al. (2013), and Koenker and Mizera (2013). Our paper may be viewed as a continuation of this effort.

The above papers extended the discussion to the situation where the observations, due to the covariates, are not exchangeable. However, the estimated parameters themselves, μ_1, \dots, μ_n , are permutation invariant. Thus, in all these problems, centering each response by a linear transformation of the covariates transforms the problem into a classical EB problem. In our setup of a time series, the estimated variables are not permutation invariant, and the explanatory variables of Y_i are the available observations Y_j , $j \neq i$, so there is an obvious strong dependence between the response variables and the covariates and the response is degenerate conditional on the covariates.

Furthermore, in all the above mentioned papers the extension of EB ideas to handle covariates is done in a retrospective setup, where all the observations are given in advance. Under the time series structure that we study, it is natural to consider real time sequential estimation of the μ 's. In Section 3 we consider the sequential case, where at stage i the decision function should be approximated based on the currently available data. Our analysis would be based on an extension of Samuel (1965). The retrospective case is simpler and will be treated first in Section 2. A small simulation study is presented in Section 4, and a real data example is discussed in Section 5.

1.2. Estimated simple coordinate-wise function

Most EB/CD solutions involve simple coordinate-wise functions. By the nature of the problem, these functions are estimated from the data, which is used symmetrically.

$$\hat{f}(X_1, \dots, X_n) = (\hat{f}(X_1; X_{(1)}, \dots, X_{(n)}), \dots, \hat{f}(X_n; X_{(1)}, \dots, X_{(n)})), \quad (4)$$

where $X_{(1)} \leq \dots \leq X_{(n)}$ are the ordered statistics

Unfortunately, any permutation invariant function can be written in this way. Suppose for simplicity that X_1, \dots, X_n are real. Let $\psi(X_1, \dots, X_n) : \mathbb{R}^n \rightarrow \mathbb{R}^n$

be a permutation invariant function. Let $\mathbf{1}(\cdot)$ be the indicator function. It is possible to write $\psi = \hat{\mathbf{f}}$ as in (4), with

$$\hat{\mathbf{f}}(x; X_{(1)}, \dots, X_{(n)}) = \sum \psi_i(X_{(1)}, \dots, X_{(n)}) \mathbf{1}(x = X_{(i)}),$$

or a smooth version of this function.

Actually, any function that is estimated from the data and is used only on that data can be written as a simple coordinate-wise function.

Intuitively, the set of simple coordinate-wise functions is a strict subset of the set of permutation invariant functions. We therefore consider a function $\hat{\mathbf{f}}$ as simple coordinate-wise function if it approximates a function \mathbf{f} that is simple coordinate-wise function by Definition 1. This later function may be random (i.e., a stochastic process), with non-degenerate asymptotic distribution.

1.3. Assumptions

The performance of our estimators will be measured by their mean squared error loss, in vector notation: $E\|\boldsymbol{\mu}^I - \boldsymbol{\mu}\|^2$, $E\|\tilde{\boldsymbol{\mu}} - \boldsymbol{\mu}\|^2$, and $E\|\hat{\boldsymbol{\mu}} - \boldsymbol{\mu}\|^2$. Let \mathcal{F}_i be the smallest σ -field under which Y_j , $j \in \mathcal{D}_i$ are measurable. The dependency of different objects on n will be suppressed, when there will be no danger of confusion.

Assumption 1 For every $i = 1, \dots, n$, the estimator $\tilde{\mu}_i$ is \mathcal{F}_i measurable. It is Lipschitz in Y_j with a constant $\rho_{|i-j|}$, where $\limsup_{M \rightarrow \infty} M^2 \rho_M < 1$. That is: For $j \in \mathcal{D}_i$ let $\tilde{\mu}'_i$ be $\tilde{\mu}_i$, but computed with Y_j replaced by $Y_j + d$. Then, $|\tilde{\mu}'_i - \tilde{\mu}_i| \leq \rho_{|i-j|}d$.

This condition is natural when $\tilde{\mu}$ is KF for a stationary Gaussian process, where typically β_{ij} decreases exponentially with $|i - j|$. The main need for generalizing the KF is to include filters which are based on estimated parameters.

The Kalman filter for an ergodic process also satisfies the following condition. It has no real importance for our results, except giving a standard benchmark.

Assumption 2 Suppose that there is a $\alpha_n \in \mathcal{F}_n$, $\alpha_n < 1$:

$$\hat{\mu}_i = \alpha_n \tilde{\mu}_i + (1 - \alpha_n)Y_i + \zeta_i, \text{ where } E\zeta_i^2 \rightarrow 0, \quad (5)$$

as $n \rightarrow \infty$, $0 < \liminf i/n \leq \limsup i/n < 1$.

Remark: Our major example is the ergodic normal state-space model. If the assumed model is correct, and $\tilde{\mu}_i$ and $\hat{\mu}_i$ are the optimal estimators, then $\tilde{\mu}_i$ is a sufficient statistics for μ_i given \mathcal{D}_i . The estimators satisfy (5) with $\alpha_n \equiv (1 + \tau^2)^{-1}$, where τ^2 is the asymptotic variance of μ_i given $\tilde{\mu}_i$. In the iterative Kalman filter method for computing $\hat{\mu}_i$, with some abuse of notation, the values $\alpha_i = (1 + \tau_i^2)^{-1}$ are computed, with τ_i^2 the variance of μ_i given $\tilde{\mu}_i$, and we have $\hat{\mu}_i = \alpha_i \tilde{\mu}_i + (1 - \alpha_i)Y_i$.

By considering the functions $g(z) \equiv 0$ and $g(z) = (1 - \alpha)z$ in (2), it is easy to see that $\boldsymbol{\mu}^I$ has asymptotically mean squared error not larger than $\tilde{\boldsymbol{\mu}}$ and $\hat{\boldsymbol{\mu}}$, respectively. In fact, we argue that unless the process is asymptotically Gaussian, there is a strict improvement.

The derivation of the Kalman filter is based on an assumed stochastic model for the sequence μ_1, \dots, μ_n . Very few properties of the the process are relevant, and it is irrelevant to our discussion whether the model is true or not. However, we do need some tightness. We expect that typically $|\mu_i - \mu_{i-1}|$ is not larger than $\log n$, and $\tilde{\mu}_i$ is sensible at least as $\tilde{\mu}_i \equiv Y_{i-1}$. Since $\max |Y_i - \mu_i| = o_p(\sqrt{\log n})$, the next condition is natural:

Assumption 3 It holds:

$$\frac{1}{n} \sum_{i=1}^n P(|Y_i - \tilde{\mu}_i| > \log n) \leq \frac{1}{(\log n)^8}.$$

2. Retrospective estimation

Denote,

$$\begin{aligned} Z_i &= Y_i - \tilde{\mu}_i; \\ \nu_i &= \mu_i - \tilde{\mu}_i, \quad i = 1, \dots, n. \end{aligned} \tag{6}$$

Clearly, $Z_i = \nu_i + \varepsilon_i$. Since ε_i is independent both of μ_1, \dots, μ_n and of $\varepsilon_j, j \neq i$, it is independent of ν_i . Thus, the conditional distribution of Z_i given ν_i is $N(\nu_i, 1)$. However, this is not a regular EB problem. It is not so even for the regular KF. Write $\tilde{\boldsymbol{\mu}} = B\mathbf{Y} = B\boldsymbol{\mu} + B\boldsymbol{\varepsilon}$. Then $\boldsymbol{\nu} = (I - B)\boldsymbol{\mu} - B\boldsymbol{\varepsilon}$. It is true that $\mathbf{Z} = \boldsymbol{\nu} + \boldsymbol{\varepsilon}$, but the vectors $\boldsymbol{\nu}$ and $\boldsymbol{\varepsilon}$ are not independent. Hence $\mathbf{Z}|\boldsymbol{\nu} \not\sim N_n(\boldsymbol{\nu}, I_n)$. Yet, we rely only on the marginal distributions of $Z_i|\nu_i, i = 1, \dots, n$.

To elaborate,

$$\begin{pmatrix} \mathbf{Y} \\ \boldsymbol{\mu} \end{pmatrix} = \begin{pmatrix} (I - B)^{-1} & 0 \\ B(I - B)^{-1} & I \end{pmatrix} \begin{pmatrix} \mathbf{Z} \\ \boldsymbol{\nu} \end{pmatrix}.$$

Therefore, the joint density of \mathbf{Z} and $\boldsymbol{\nu}$ is proportional to

$$f_{\boldsymbol{\mu}}(\boldsymbol{\nu} + B(I - B)^{-1}\mathbf{Z}) \exp(-\|\mathbf{Z} - \boldsymbol{\nu}\|^2/2),$$

where $f_{\boldsymbol{\mu}}$ is the joint density of the vector $\boldsymbol{\mu}$. Clearly, unless $f_{\boldsymbol{\mu}}$ is multivariate normal, the conditional density of \mathbf{Z} given $\boldsymbol{\nu}$ is not multivariate standard normal.

Example 2.1 Suppose $n = 2$, we observe Y_0, Y_1 , and use $\tilde{\mu}_i = \gamma Y_{1-i}, i = 0, 1$. Then

$$\begin{aligned} Z_i &= Y_i - \gamma Y_{1-i} &\Rightarrow Y_i &= \frac{1}{1 - \gamma^2} (Z_i + \gamma Z_{1-i}) \\ \nu_i &= \mu_i - \gamma Y_{1-i} &\Rightarrow Y_i &= \frac{1}{\gamma} (\mu_{1-i} - \nu_{1-i}) \end{aligned}$$

$$\Rightarrow Z_i = \frac{1}{\gamma}(\mu_{1-i} - \nu_{1-i} - \gamma\mu_i + \gamma\nu_i).$$

Suppose further that μ_i is finitely supported. It follows from the above calculations that the distribution of the vector \mathbf{Z} given the vector $\boldsymbol{\nu}$ is finitely supported as well.

The estimator in vector notation is $\boldsymbol{\mu}^I = \tilde{\boldsymbol{\mu}} + \boldsymbol{\delta}$, where $\boldsymbol{\delta} = (\delta(Z_1), \dots, \delta(Z_n))^\top$. As discussed in the introduction, simple coordinate-wise functions like $\boldsymbol{\delta}$ are central in EB and CD models. However, our decision function $\boldsymbol{\mu}^I$ is not a simple coordinate-wise function of the observations. It is a hybrid of non-coordinate-wise function $\tilde{\boldsymbol{\mu}}$ and a simple coordinate-wise one, $\boldsymbol{\delta}$. The $\tilde{\boldsymbol{\mu}}$ component accounts for the non-coordinate-wise information from the covariates, while $\boldsymbol{\delta}$ aims to improve it in a coordinate-wise way after the information from all other observations was accounted for by $\tilde{\boldsymbol{\mu}}$.

By Assumption 1, the dependency between the Z_i s conditioned on $\boldsymbol{\nu}$ is only local, and hence, if we consider a permutation invariant procedure, which treats neighboring observations and far away ones the same, the dependency disappears asymptotically, and we may consider only the marginal normality of the $\boldsymbol{\nu}$. The basic ideas of EB are helpful and we get the representation (7) of $\boldsymbol{\delta}$ as given below.

Let

$$f_Z(z) = \frac{1}{n} \sum_{i=1}^n \varphi(z - \nu_i),$$

where φ is the standard normal density. Note that this is not a kernel estimator—the kernel is with fixed bandwidth and ν_1, \dots, ν_n are unobserved. Let I be uniformly distributed over $1, \dots, n$. Denote by F^n the distribution of the random pairs $(\nu_I, \nu_I + \eta_I)$, where η_1, \dots, η_n are i.i.d. standard normal independent of the other random variables mentioned so far and the randomness is induced by the random index I and the η s. One marginal distribution of F^n is the empirical distribution of ν_1, \dots, ν_n , while the density of the other is given by f_Z . We denote the marginals by F_ν^n and F_Z^n . Finally, note that Z_I given ν_I has the distributing of $F_{Z|\nu}^n$, i.e., the conditional distribution of F^n .

It is well known that asymptotically, the Bayes procedure for estimating ν_i given Z_i is approximated by the Bayes procedure with F_ν^n as prior, and it is determined by f_Z . The optimal simple coordinate-wise function $\delta = \delta^n$ depends only on marginal joint distribution of (ν_I, Z_I) . In fact, it depends only on f_Z . As in Brown (1971) we have:

$$\delta^n(z) = E_{F^n}(\nu_I | \nu_I + \eta_I = z) = z + \frac{f_Z'(z)}{f_Z(z)}, \quad (7)$$

where f_Z' is the derivative of f_Z . The dependency on n is suppressed in the notations.

Note that δ^n is a random function, and in fact, if μ_1, \dots, μ_n is not an ergodic process, it may not have an asymptotic deterministic limit. Yet, it would be the object we estimate in (8) below.

It is of a special interest to characterize when $\delta = \delta^n$ is asymptotically linear, in which case the improved estimator $\mu_i^I = \tilde{\mu}_i + \delta^n(Z_i)$ is asymptotically a linear combination of $\tilde{\mu}_i$ and Y_i . Only in such a case the difference between the loss of the improved estimator $\mu^I = \tilde{\mu} + \delta$ and that of the estimator $\hat{\mu}$ may be asymptotically of $o(n)$. It follows from (7) that, the optimal decision $\delta(Z)$ is approximately $(1 - \alpha)Z$, if and only if, $f'_Z/f_Z = (\log f_Z)'$ is approximately proportional to z . This happens only if f_Z converges to a Gaussian distribution. Since f_Z is a convolution of a Gaussian kernel with the prior, this can happen only if the prior is asymptotically Gaussian. In our setup where F_ν^n plays the role of a prior, in order to have asymptotically linear improved estimator we need that F_ν^n converges weakly to a normal distribution G .

The above is formally stated in the following Proposition 2.1.

Proposition 2.1 *Under assumptions 1-3, $n^{-1}E\|\mu^I - \hat{\mu}\|^2 \rightarrow 0$ implies that the sequence $F_\nu^n - N(0, (1 - \alpha_n)/\alpha_n)$ converges weakly to the zero measure.*

Given the observations Z_1, \dots, Z_n , let \hat{F}_Z^n be the empirical distribution of Z_1, \dots, Z_n . We will show that as $n \rightarrow \infty$ the ‘distance’ between \hat{F}_Z^n and F_Z^n gets smaller, so that f_Z and its derivative may be replaced in (7) by appropriate kernel estimates based on \hat{F}_Z^n , and yield a good enough estimator $\hat{\delta}^n$ of δ^n . In the sequel we will occasionally drop the superscript n . Note, that we do not assume that F^n approaches some limit F as $n \rightarrow \infty$, although this is the situation if we assume that Z_1, Z_2, \dots is an ergodic stationary process, however our assumptions on that process are milder.

We now state the above formally. Consider the two kernel estimators $\hat{f}_Z(z) = n^{-1} \sum_j K_\sigma(Z_j - z)$ and $\hat{f}'_Z(z) = n^{-1} \sum_j K'_\sigma(Z_j - z)$, where $K_\sigma(z) = \sigma^{-1}K(z/\sigma)$, $\sigma = \sigma_n$. For simplicity, we use the same bandwidth to estimate both the density and its derivative. We define the following estimator $\hat{\delta} \equiv \hat{\delta}_\sigma$ for δ :

$$\hat{\delta}(z) = \hat{\delta}_\sigma^n(z) \equiv z + \frac{\hat{f}'_Z(z)}{\hat{f}_Z(z)}. \quad (8)$$

Brown and Greenstein (2009) used the normal kernel, we prefer to use the logistic kernel $2(e^x + e^{-x})^{-1}$ (it is the derivative of the logistic cdf, $(1 + e^{-2x})^{-1}$, and hence its integral is 1). We suggest this kernel since it ensures that $|\hat{\delta}(z) - z| < \sigma^{-1}$, see the Appendix. However, we do adopt the recommendation of Brown and Greenshtein (2009) for a very slowly converging sequence $\sigma_n = 1/\log(n)$.

Denote $\hat{\mu}^I = \tilde{\mu} + \hat{\delta}$.

Theorem 2.2 *Under Assumptions 1 and 3:*

i)

$$E\|\hat{\mu}^I - \mu\|^2 \leq E\|\mu^I - \mu\|^2 + o(n) \leq E\|\hat{\mu} - \mu\|^2 + o(n).$$

ii) *Under Assumptions 1-3, if $\alpha_n \xrightarrow{P} \alpha \in (0, 1)$ and F_ν^n converges weakly to a distribution different from $N(0, (1 - \alpha)/\alpha)$, then there is $c < 1$ such that*

for large enough n :

$$E\|\hat{\boldsymbol{\mu}}^I - \boldsymbol{\mu}\|^2 \leq cE\|\hat{\boldsymbol{\mu}} - \boldsymbol{\mu}\|^2.$$

Proof.

- i) The proof is given in the appendix
- ii) The same arguments used to prove part i) may be used to prove a modification of Proposition 2.1, in which μ_i^I is replaced by $\hat{\mu}_i^I$, when assuming in addition that $i/n \in (\zeta, 1 - \zeta)$ for any $\zeta \in (0, 1)$.

□

Part i) of the above theorem assures us that asymptotically the improved estimator does as good as $\hat{\boldsymbol{\mu}}$; part ii) implies that in general the improved estimator does asymptotically strictly better. Obviously the asymptotic improvement is not always strict since the Kalman filter is optimal under a Gaussian state-space model.

3. Sequential estimation

We consider now the case $\mathcal{F}_i = \sigma(Y_1, \dots, Y_{i-1})$, $i \leq n$. The definition of the different estimators is the same as in the previous section with the necessary adaption to the different information set. Our aim is to find a sequential estimator, denoted $\hat{\boldsymbol{\mu}}^{IS}$, that satisfies $E\|\hat{\boldsymbol{\mu}}^{IS} - \boldsymbol{\mu}\|^2 + o(n) < E\|\hat{\boldsymbol{\mu}} - \boldsymbol{\mu}\|^2$. By a sequential estimator $\hat{\boldsymbol{\mu}}^{IS} = (\psi^1, \dots, \psi^n)$ we mean that $\psi^i \in \mathcal{F}_i$, $i = 1, \dots, n$. A natural approach, which indeed works, is to let $\psi^i = \tilde{\mu}_i + \hat{\delta}^i$, where $\hat{\delta}^i$ is defined as in (8), but with $\hat{f} = \hat{f}_i$ restricted to the available data Z_1, \dots, Z_{i-1} , $i = 1, \dots, n$. Let $\hat{\boldsymbol{\delta}}^S = (\hat{\delta}^1, \dots, \hat{\delta}^n)$.

We define:

$$\hat{\boldsymbol{\mu}}^{IS} = \tilde{\boldsymbol{\mu}} + \hat{\boldsymbol{\delta}}^S.$$

Our main result in this section:

Theorem 3.1 *Theorem 2.2 holds with $\hat{\boldsymbol{\mu}}^{IS}$ and $\boldsymbol{\mu}^{IS}$ replacing $\hat{\boldsymbol{\mu}}^I$ and $\boldsymbol{\mu}^I$, respectively.*

In order to prove Theorem 3.1 we adapt Lemma 1 of Samuel (1965). Samuel's result is stated for a compound decision problem, i.e., the parameters are fixed, and the observations are independent. The result compares the performance of the optimal estimators in the sequential and retrospective procedures. It is not clear a priori whether retrospective estimation is easier or more difficult than the sequential. On the one hand, the retrospective procedure is using more information when dealing with the i th parameter. On the other hand, the sequential estimator can adapt better to non-stationarity in the parameter sequence. Samuel proved that the latter is more important. There is no paradox here, since the retrospective procedure is optimal only under the assumption of permutation invariance, and under permutation invariance, the weak inequality in Lemma 3.2 below is, in fact, equality.

Our approach is to rephrase and generalize Samuel's lemma. Let η_1, \dots, η_n be $N(0, 1)$ i.i.d. random variables independent of (μ_i, ε_i) , $i = 1, \dots, n$. Let $L(\nu_i, \hat{\nu}_i)$ be the loss for estimating ν_i by $\hat{\nu}_i$. For every $i \leq n$ let δ^i be the decision function that satisfies:

$$\begin{aligned} \delta^i &= \arg \min_{\delta} E_{\boldsymbol{\eta}} \sum_{j=1}^i L(\nu_j, \delta(\nu_i + \eta_i)) \\ &= \arg \min_{\delta} \sum_{j=1}^i E_{\boldsymbol{\eta}} L(\nu_j, \delta(\nu_i + \eta_i)) \\ &\equiv \arg \min_{\delta} \sum_{j=1}^i R(\delta, \nu_j), \quad \text{say,} \end{aligned}$$

where $E_{\boldsymbol{\eta}}$ is the expectation over $\boldsymbol{\eta}$. That is, δ^i is the functional that minimizes the sum of risks for estimating the components ν_1, \dots, ν_i , but it is applied only for estimating ν_i . The quantity $R(\delta^j, \nu_j)$ is the analog of $R(\phi_{F_j}, \theta_j)$ in Samuel's formulation. In analogy to Samuel (1965) we define $R_n \equiv n^{-1} \sum_{j=1}^n R(\delta^n, \nu_j)$, the empirical Bayes risk of the non-sequential problem.

Lemma 3.2

$$n^{-1} \sum_{j=1}^n R(\delta^j, \nu_j) \leq R_n.$$

The proof of the lemma is formally similar to the proof of Lemma 1 of Samuel (1965).

Proof of Theorem 3.1. From Theorem 2.2, $E(\hat{\delta}^i(Z) - \delta^i(Z))^2 \rightarrow 0$. The last fact coupled with Lemma 3.2 implies part i) of Theorem 3.1. Part ii) is shown similarly to part ii) of Theorem 2.2. \square

4. Simulations.

We present now simulation results for the following state-space model.

$$\begin{aligned} Y_i &= \mu_i + \varepsilon_i \\ \mu_i &= \phi \mu_{i-1} + U_i, \quad i = 1, \dots, n, \end{aligned} \tag{9}$$

where $\varepsilon_i \sim N(0, 1)$, $i = 1, \dots, n$, are independent of each other and of U_i , $i = 1, \dots, n$. The variables U_i , $i = 1, \dots, n$ are independent, $U_i = X_i I_i$ where $X_i \sim N(0, v)$ are independent, while I_1, \dots, I_n are i.i.d. Bernoulli with mean 0.1, independent of each other and of X_i , $i = 1, \dots, n$. We study the twelve cases that are determined by $\phi = 0.25, 0.75$ and $v = 0, 1, \dots, 5$. In each case we investigate both the sequential and the retrospective setups.

If U_1, \dots, U_n , were i.i.d Normal, the data would follow a Gaussian state-space, and the corresponding Kalman filter estimator would be optimal. Since the U_i 's are not normal, the corresponding AR(1) Kalman filter is not optimal (except in the degenerate case, $v = 0$), though it is optimal among linear filters. This is reflected in our simulation results where for the cases $v = 0, 1$ our “improved” method $\hat{\mu}^I$ performs slightly worse than $\hat{\mu} = \hat{\mu}^K$. It improves in all the rest. The above is stated and proved formally in the following proposition. It could also be shown indirectly by applying part ii) of Theorem 2.2.

Proposition 4.1 *Consider the state-space model, as defined by (9). If U_i are not normally distributed then*

$$E\|\hat{\mu}^I - \mu\|^2 \leq cE\|\hat{\mu}^K - \mu\|^2,$$

for a constant $c \in (0, 1)$ and large enough n .

Proof. Given the estimators $\tilde{\mu}_i^K$ and $\hat{\mu}_i^K$, $i = 1, \dots, n$, let $Z_i = Y_i - \tilde{\mu}_i$. Then $Z_i = \mu_{i-1} + U_i - \tilde{\mu}_i^K + \varepsilon_i = \nu_i + \varepsilon_i$. The distribution G^i of ν_i may be normal only if U_i is normal, since U_i is independent of μ_{i-1} and $\tilde{\mu}_i$. The distributions G^i converge to a distribution G as i and $n - i$ approach infinity. As before, G is normal only if U_i are normal. Now, asymptotically optimal estimator for μ_i under squared loss and given the observation Y_i , is $\tilde{\mu}_i^K + \hat{\nu}_i$, where $\hat{\nu}_i$ is the Bayes estimator under a prior G on ν_i and an observation $Y_i \sim N(\nu_i, 1)$. This Bayes estimator is linear and coincide with the KF estimator $\hat{\mu}_i^K$, only if G is normal. □

Analogous discussion and situation are valid also in the sequential case. In our simulations the parameters ϕ and $VAR(U_i)$ are treated as known. Alternatively, maximum likelihood estimation assuming (wrongly) normal inovations yields results similar to those reported in Table 1.

The simulation results in Table 1 are for the case $n = 500$. Each entry is based on 100 simulations. In each realization we recorded $\|\hat{\mu} - \mu\|^2$ and $\|\hat{\mu}^I - \mu\|^2$, and each entry is based on the corresponding average. In order to speed the asymptotics we allowed a ‘warm up’ of 100 observations prior to the $n = 500$ in the sequential case, we also allowed a ‘warm up’ of 50 in both sides of the $n = 500$ observations in the retrospective case.

It may be seen that when the best linear filter is optimal or nearly optimal (when $v = 0$ or approximately so), our improved method is slightly worse than the Kalman filter estimator, however as v increases, the advantage of the improved method may become significant.

It seems that in the case $\phi = 0.25$ the future observations are not very helpful, and in our simulations there are cases where the simulated risk of $\hat{\mu}$ in the sequential case is even smaller than the corresponding simulated risk of the retrospective case. This could be an artifact of the simulations, but also a result of noisy estimation of the coefficients β_{ij} of the mildly informative future observations.

TABLE 1
Mean Squared error of the two estimator for an autoregressive process with aperiodic normal shocks

ϕ	0.25						0.75					
v	0	1	2	3	4	5	0	1	2	3	4	5
<i>Retrospective filter:</i>												
$\hat{\mu}^\dagger$	0	71	156	226	290	333	0	49	147	235	301	350
$\mu^{I\dagger}$	23	66	125	148	160	177	24	91	166	215	253	271
<i>Sequential filter:</i>												
$\hat{\mu}^\dagger$	0	47	145	234	309	355	0	83	187	264	325	372
$\mu^{I\dagger}$	39	81	129	147	159	158	34	112	184	216	239	253

\dagger Kalman filter, \ddagger Improved.

TABLE 2
The retrospective case: Cross-validation estimation of the average squared risk.

$p = 0.95$	AR(1)	AR(2)	ARIMA(1,1,0)
Kalman filter - $\hat{\lambda}_i^K$	27.1	19.4	20.4
Improved method - $\hat{\lambda}_i^I$	18.7	18.5	17.4
Naive method - $\hat{\lambda}_i^N$	26.4	26.4	26.4

5. Real Data Example

In this section we demonstrate the performance of our method on real data taken from the FX (foreign exchange) market in Israel. The data consists of the daily number of swaps (purchase of one currency for another with a given value date, done simultaneously with the selling back of the same amount with a different value date. This way there is no foreign exchange risk) in the OTC (over-the-counter) Shekel/Dollar market. We consider only the buys of magnitude 5 to 20 million dollars. The time period is January 2nd, 2009 to December 31st, 2013, a total of $n = 989$ business days. The number of buys in each day is 24 on the average, with the range of 2–71. In our analysis we used the first 100 observations as a ‘warm up’, similarly to the way it was done in our simulations section.

We denote by X_i , $i = 1, \dots, n$, the number of buys on day i and assume that $X_i \sim Po(\lambda_i)$. We transform the data by $Y_i = 2\sqrt{X_i} + 0.25$ as in Brown et al. (2010) and Brown et al. (2013) in order to get an (approximately) normal variable with variance $\sigma^2 = 1$.

The assumed model in this section is the following state space system of equations:

$$Y_i = \mu_i + \varepsilon_i$$

$$\mu_i \sim ARIMA(p, d, q), \quad i = 1, \dots, n,$$

where $\mu_i = 2\sqrt{\lambda_i}$ and $\varepsilon_i \sim N(0, 1)$ are independent of each other and of the $ARIMA(p, d, q)$ process. We consider the following three special cases of $ARIMA(p, d, q)$: $AR(1)$, $AR(2)$, and $ARIMA(1, 1, 0)$.

Under each model there are induced Kalman filter estimators $\tilde{\mu}^K$, and $\hat{\mu}^K$ that correspond to one step prediction and to the update. Similarly, the im-

proved estimator $\hat{\mu}^I$ is defined. We denote the sequential and retrospective estimators similarly with no danger of confusion.

After estimating μ_i , we transform the result back to get the estimator $\hat{\lambda}_i^J$ for λ_i , $\hat{\lambda}_i^J = 0.25 \left(\hat{\mu}_i^J \right)^2$, $i = 1, \dots, n$, $J \in \{ 'I', 'K' \}$ where, $\hat{\mu}_i^J$ is the estimator of μ_i by method J . We evaluate the performances of both estimation methods by the following non-standard cross-validation method as described in Brown et al. (2013). It is briefly explained in the following.

Let $p \in (0, 1)$, $p \approx 1$, and let U_1, \dots, U_n be independent given X_1, \dots, X_n , where $U_i \sim B(X_i, p)$ are Binomial variables. It is known that $U_i \sim Po(p\lambda_i)$ and $V_i = X_i - U_i \sim Po((1-p)\lambda_i)$ and they are independent given $\lambda_1, \dots, \lambda_n$. We will use the ‘main’ sub-sample U_1, \dots, U_n for the construction of both estimators (Kalman filter and Improvement) while the ‘auxiliary’ sub-sample V_1, \dots, V_n is used for validation. Consider the following loss function,

$$\begin{aligned} \rho(J; \mathbf{U}, \mathbf{V}) &= \frac{1}{n} \sum_{i=1}^n \left(\frac{\hat{\lambda}_i^J}{p} - \frac{V_i}{(1-p)} \right)^2 \\ &= \frac{1}{np^2} \sum_{i=1}^n \left(\hat{\lambda}_i^J - p\lambda_i \right)^2 + \frac{1}{n(1-p)^2} \sum_{i=1}^n (V_i - (1-p)\lambda_i)^2 \\ &\quad - \frac{2}{n} \sum_{i=1}^n \left(\frac{\hat{\lambda}_i^J}{p} - \lambda_i \right) \left(\frac{V_i}{(1-p)} - \lambda_i \right) \\ &= \frac{1}{np^2} \sum_{i=1}^n \left(\hat{\lambda}_i^J - p\lambda_i \right)^2 + A_n + R_n(J), \quad J \in \{ 'K', 'I' \}. \end{aligned}$$

The term $R_n(J) = \mathcal{O}_p(n^{-1/2})$ and will be ignored. We estimate A_n by the method of moments:

$$\hat{A}_n = \frac{1}{n(1-p)^2} \sum_{i=1}^n V_i.$$

We repeat the cross-validation process 500 times and average the computed values of $\rho(J; \mathbf{U}, \mathbf{V}) - \hat{A}_n$. When p is close to 1, the average obtained is a plausible approximation of the average squared risk in estimating λ_i , $i = 101, \dots, 989$. By the above method we approximated also the average risk of the naive estimator $\hat{\lambda}_i^N = X_i$, $i = 101, \dots, 989$. The approximations for the retrospective and sequential cases, are displayed in Tables 2 and 4. The estimated ARIMA coefficient for the various models are given in Table 3.

From Table 2 we can observe that in the retrospective case the improved method does uniformly better than the naive estimator and the Kalman filter. In fact, in all except a small deterioration under the ARIMA(1,1,0) with sequential filtering, the performance of the improved method is quite uniform, showing its robustness against model miss-specification.

Somewhat surprising is that the Kalman filter under AR(1) with retrospective estimation does not do better than the naive filter, but do so considerably in

TABLE 3
The retrospective case: Parameter estimation

$p = 0.95$	AR(1)	AR(2)	ARIMA(1,1,0)
α	12.38	14.218	0.01
ϕ_1	-0.28	-0.341	-0.6
ϕ_2		-0.124	
σ^2	3.4	3.4	4.7

TABLE 4
The sequential case: Cross-validation approximation of the average squared risk

$p = 0.95$	AR(1)	AR(2)	ARIMA(1,1,0)
Kalman filter - $\hat{\lambda}_i^K$	19.2	19.2	21.2
Improved method - $\hat{\lambda}_i^I$	19.0	19.2	22.6
Naive method - $\hat{\lambda}_i^N$	26.4	26.4	26.4

the sequential case. The reason is that the AR(1) model does not fit the data. When it is enforced on the data, the Kalman filter gives too much weight to the surrounding data, and too little to the “model free” naive estimator. This result show the robustness of our estimator.

In fact, we did a small simulation, where the process was AR(2), with the parameters as estimated for the data. When an AR(1) was fitted to the data, the retrospective Kalman filter was strictly inferior to the sequential one.

In the sequential case, Table 4, the improved method does better than the naive method, but contrary to the non-sequential case, it improves upon the Kalman filter only in the AR(1) and AR(2) models, while in the ARIMA(1,1,0) model the Kalman filter does slightly better.

6. Appendix: Proof of Theorem 2.2

Note that by (3), obviously, $E\|\boldsymbol{\mu}^I - \boldsymbol{\mu}\|^2 < E\|\hat{\boldsymbol{\mu}} - \boldsymbol{\mu}\|^2 + o(n)$. Thus, in order to obtain $E\|\hat{\boldsymbol{\mu}}^I - \boldsymbol{\mu}\|^2 < E\|\hat{\boldsymbol{\mu}} - \boldsymbol{\mu}\|^2 + o(n)$ it is enough to show that $E\|\hat{\boldsymbol{\delta}} - \boldsymbol{\delta}\|^2 = o(n)$.

First recall:

$$\begin{aligned}
 K(x) &= \frac{2}{(e^x + e^{-x})^2} \\
 K'(x) &= -4 \frac{e^x - e^{-x}}{(e^x + e^{-x})^3} \\
 \frac{K'(x)}{K(x)} &= -2 \frac{e^x - e^{-x}}{e^x + e^{-x}} \in (-2, 2).
 \end{aligned}$$

Thus

$$\sup_z \left| \frac{\hat{f}'_Z(z)}{\hat{f}_Z(z)} \right| < 2\sigma_n^{-1}. \quad (10)$$

By Assumption 1, if we replace $\tilde{\mu}_i$ by a similar (unobserved) function $\tilde{\mu}_i^*$, where Y_j , $j \in \mathcal{D}_i$, $|j - i| > n^\gamma$ is replaced by μ_j , then $\max_i |\tilde{\mu}_i^* - \tilde{\mu}_i| =$

$\mathcal{O}_p(n^{-\gamma}) \max_i |\varepsilon_i| = \mathcal{O}_p(n^{-\gamma/2})$, for any $\gamma > 0$. Define ν_i^* and Z_i^* as in (6), but where $\tilde{\mu}_i$ is replaced by $\tilde{\mu}_i^*$, $i = 1, \dots, n$. Since $|\hat{f}''| \leq \sigma_n^{-3}$, $\max_i |Z_i - Z_i^*|$ is ignorable for our approximations. In the following all variables are replaced by their $*$ version, but we drop the $*$ for simplicity.

Let $L_n = \log n$. By Assumption 3 with probability greater than $1 - L_n^{-4}$, all but n/L_n^4 of the Z_i s are in $(-L_n, L_n)$, hence, their density is mostly not too small:

$$\int \mathbf{1}(f_Z(z) < L_n^{-3}) f_Z(z) dz < L_n^{-2}. \quad (11)$$

Let

$$\begin{aligned} \bar{\varphi}(z) &= \varphi * K_{\sigma_n} = \int K\left(\frac{z-e}{\sigma_n}\right) \varphi(e) de \\ \bar{f}_Z &= f_Z * K_{\sigma_n} = \frac{1}{n} \sum_{j=1}^n \bar{\varphi}(z - \nu_j) \end{aligned}$$

Then

$$\begin{aligned} \hat{f}_Z(z) - \bar{f}_Z(z) &= \frac{1}{\sigma_n n} \sum_{j=1}^n K\left(\frac{z - Z_j}{\sigma_n}\right) - \frac{1}{n} \sum_{j=1}^n \bar{\varphi}(z - \nu_j) \\ &= \frac{1}{\sigma_n n} \sum_{j=1}^n \left\{ K\left(\frac{z - \nu_j - \varepsilon_j}{\sigma_n}\right) - \int K\left(\frac{z - \nu_j - e}{\sigma_n}\right) \varphi(e) de \right\}. \end{aligned}$$

Since ε_j and ν_j are independent, this difference has mean 0. Moreover, since (ν_j^*, ε_j) and (ν_k^*, ε_k) are independent for $|j - k| > M$, the above expression has variance of order $M(n\sigma_n)^{-1}$. The difference $\|\bar{f}_Z - f_Z\|_\infty$ is of order σ_n^2 .

A similar expansion works for \hat{f}'_Z , except that the variance now is of order $M(n\sigma_n^3)^{-1}$. Regular large deviation argument shows that when $f_Z(Z_i) > L_n^{-3}$ then $\hat{f}_Z(Z_i) < L_n^{-3}/2$ with exponential small probability. By (10) and (11) it follows that the approximation $\hat{f}'_Z(Z_i)/\hat{f}_Z(Z_i) = f'_Z(Z_i)/f_Z(Z_i) + \mathcal{O}_p(1)$ holds not only in the mean but also in the mean square, since the exception probability are smaller than σ_n^2 .

References

- Brockwell, P.J. and Davis, R.A. (1991). Time Series: Theory and Methods. Second edition. Springer.f
- Brown, L. D. (1971). Admissible estimators, recurrent diffusions, and insoluble boundary value problems. *Ann.Math.Stat.* **42**, 855-904.
- Brown, L.D., Cai, T., Zhang, R., Zhao, L., Zhou, H. (2010). The root-unroot algorithm for density estimation as implemented via wavelet block thresholding. *Probability and Related Fields*, **146**, 401-433.
- Brown, L.D. and Greenshtein, E. (2009). Non parametric empirical Bayes and compound decision approaches to estimation of high dimensional vector of normal means. *Ann. Stat.* **37**, No 4, 1685-1704.

- Brown L.D., Greenshtein, E. and Ritov, Y. (2013). The Poisson compound decision revisited. *JASA*. **108** 741-749.
- Cohen, N., Greenshtein E., and Ritov, Y. (2012). Empirical Bayes in the presence of explanatory variables. *Statistica Sinica*. **23**, No. 1, 333-357.
- Copas, J.B. (1969). Compound decisions and empirical Bayes (with discussion). *JRSSB* **31** 397-425.
- Fay, R.E. and Herriot, R. (1979). Estimates of income for small places: An application of James-Stein procedure to census data. *JASA*, **74**, No. 366, 269-277.
- Greenshtein, E. and Ritov, Y. (2008). Asymptotic efficiency of simple decisions for the compound decision problem. The 3'rd Lehmann Symposium. IMS Lecture Notes Monograph Series, J.Rojo, editor. 266-275.
- Jiang, W. and Zhang, C.-H. (2010). Empirical Bayes in-season prediction of baseball batting average. *Borrowing Strength: Theory Powering Application-A festschrift for L.D. Brown* J.O. Berger, T.T. Cai, I.M. Johnstone, eds. IMS collections **6**, 263-273.
- Koenker, R. and Mizera, I. (2012). Shape constraints , compound decisions and empirical Bayes rules. Manuscript.
- Robbins, H. (1951). Asymptotically subminimax solutions of compound decision problems. *Proc. Second Berkeley Symp.* 131-148.
- Robbins, H. (1955). An Empirical Bayes approach to statistics. *Proc. Third Berkeley Symp.* 157-164.
- Robbins, H. (1964). The empirical Bayes approach to statistical decision problems. *Ann.Math.Stat.* **35**, 1-20.
- Samuel, E. (1965). Sequential Compound Estimators. *Ann.Math. Stat.* **36**, No 3, 879-889.
- Zhang, C.-H.(2003). Compound decision theory and empirical Bayes methods.(invited paper). *Ann. Stat.* **31** 379-390.